Physiological Sensors for Speech Recognition

Mike Scanlon, Francis Fisher, Steve Chen

Abstract. *Systems designers are expressing greater interest in speech-based user interfaces for a variety of civilian and military applications. Such interfaces provide hands-free operation and a more natural way for humans to interact with systems. One difficulty with speech-based user interfaces is poor operation in noisy environments such as military operations. The Physiological Sensor, developed at ARL, is an example of an alternative sensor for automatic speech recognition. This sensor detects speech by measuring acoustic signals through the speaker's skin. While the signal produced is not typical of that from an airborne acoustic microphone, the possibility exists for using this sensor as a microphone. We investigate several possible methods for using the Physiological Sensor as a microphone for automatic speech recognition.*

1. Introduction

With recent advances in automatic speech recognition (ASR) technology has come an increased interest in applying this technology to the design of user interfaces. For a system being operated in a benign environment such an interface can be based on commercial or custom software and an airborne acoustic microphone. However, most systems of this type are difficult or impossible to use in noisy environments such as those presented in military or industrial scenarios. In such cases we must find alternative ASR software or speech sensors in order to enhance operation in these environments. Efforts to improve operation in noisy environments by removing the noise from the microphone output have proven difficult without knowledge of the external noise source.

The Physiological Sensor, a medical sensor developed at Army Research Laboratory, is a device that physically couples to a patient to record medical information such as respiration and heartbeat. With some slight modifications to the electronics, ARL has converted this sensor to a microphone to be worn around the throat.

2. Physiological Sensor - Background

ARL has developed a new method to measure human physiology and monitor health and performance parameters. This consists of an acoustic sensor positioned inside a fluid-filled bladder in contact with the human body. Packaging the sensor in this manner minimizes outside environmental interferences, and signals within the body are transmitted to the sensor bladder with minimal losses. This fluid-coupling technology comfortably conforms to the human body, and enhances the signal-to-noise-ratio (SNR) of human physiology to that of ambient noise. An acoustic sensor system can detect changes in a person's physiological status resulting from exertion or injuries such as trauma, penetrating wound, hypothermia, dehydration, heat stress, and many other conditions (or illnesses). Furthermore, a sensor contacting the torso, head, or throat region picks up the wearer's voice very well through the flesh, with fidelity sufficient to be used as an auxiliary microphone for communications or hands-free voice activation mechanism. Automatic speech recognition software, in conjunction with this enhanced body-coupling sensor, could improve mission performance by reducing false voice commands through improved SNR in noisy environments. Civilian technology transfer applications include clinical surveillance, medical transport, hospitals, and telemedicine applications. Fire, rescue, and police personnel may benefit from hands free voice communications with embedded health and performance monitoring [Scanlon, patents].

2.1 Sensor Description

The neck-band sensors shown in figures 1 and 2 consists of a housing, gel-coupling sack with sensor embedded within, neck strap, preamplifier, and battery pack with hardwired signal egress and push to talk button.. The headband sensor in figure 3 does not use a liquid coupling, but rather an acoustically conductive silicone rubber.

Data were collected at the side of the neck using a sensor of similar geometry to the sensor in figure 1 [Scanlon, 1998]. The test included a spoken word count from 1 to 10, then mouth breathing for the remainder of the data set. Naturally, the heartbeat is always present. The time and frequency representations are shown in figure 4. Figure 5 compares data from a B&K microphone in front of the speaker's mouth to that of a fluid-coupled physiological sensor held in contact with the neck by a strap. Data from both locations were taken simultaneously in a typical office environment. Comparing the amplitudes of the voice to the non-vocal ambient noise surrounding the voice gives approximately 40 dB SNR for the B&K airborne microphone, and approximately 75 dB SNR for the fluid-coupled sensor. The fluid coupling represents an

*Previously published in the Computer Science Handbook for Displays, Editors: Thomas Huang, Marius Vassiliou improvement of better than 30 dB in speech SNR with minimal waveform degradation, as observed by the similarity of spectrograms and by listening to the data through headphones.

Time (s)

Time (s)

The ability of body-coupled sensors to detect physiology and reduce background noise was investigated. A physiological sensor was attached to one side of a speaker's neck, and an omnidirectional electret microphone was place in front of the mouth. Figures 6 and 7 show simultaneously collected breath and voice data before, during, and after a speaking subject is immersed in a C-weighted noise field of 105 dB (referenced to 20 micropascals) noise field. The person wearing the sensors repeatedly vocalized a 1 to 10 count between the times of 14- and 19-s, 25- to 33-s, 65- to 71-s, and 71- to 77-s, and vocalized "105 dB" between 47- and 50-s.

The boom microphone in figure 6 does not detect any voice during the high amplitude noise between 20- and 71-s. However, in figure 7, the counting is clearly visible throughout the loud noise with the body-coupled gel sensor. Playing the data collected through headsets, the listener could clearly hear and understand the spoken words from the gel sensor in 105 dB noise, but could not discern the presence of any speech in the boom microphone data.

3. Automatic Speech Recognition Using the Physiological Sensor

Army Research Laboratory (ARL) and Rockwell Sciences Center (RSC) have developed several experimental systems that use the Physiological Sensor as input to automatic speech recognition (ASR) systems. These efforts are discussed below.

3.1 RSC Integration & application of the Physiological Sensor

3.1.1 General Signal Characteristics of the Physiological Sensor

By coupling directly to the user's neck, the physiological sensor was able to achieve extraordinary signal to noise performance as compared to airborne acoustic microphone technologies. While providing significant rejection of ambient noise, the sensor was not entirely immune to ambient sound. For instance, it was quite possible to detect other persons speaking to the wearer of the physiological sensor, though at greatly attenuated levels. Due largely to the method of transduction, the output signal of the ARL physiological sensor was significantly different from typical acoustic microphone signals. Specifically, higher frequencies tended to be significantly attenuated. Human listeners listening to the output signal of the physiological sensor indicated that the distortion was analogous to listening to a person in another room through a wall.

3.1.2 Physiological Sensor and Speech Recognizers

Because of the inherent distortions of speech associated with the ARL physiological sensor, many commercial, off-the-shelf ASR technologies, like IBM's ViaVoice, were unable to successfully recognize speech using the physiological sensor signals. Such recognizers often rely on Hidden-Markov Models of speech, where the models are pre-estimated using statistical methods and large databases of human speech. Such databases would have been collected with conventional airborne acoustic microphones, so any speaker-independent speech recognizer would have an inherent expectation about the signal characteristics of speech as normally acquired through airborne acoustic microphones. Hence, in performing speech recognition with the physiological sensor, speaker-dependent recognizers tended to work more reliably. As recommended by ARL, the initial speech recognition engine utilized was the Clamor engine, a dynamic-time-warping speech recognizer developed by the Lexicus business unit of Motorola. Clamor recorded templates of each word or phrase ("token") to be recognized as provided by the user (2 instances of each token were kept as matching templates). Performance with the Clamor recognition engine was adequate for discrete, speaker dependent recognition of up to several distinct tokens.

Later, Rockwell Science Center developed a speaker-dependent, Hidden-Markov Model based discrete speech recognizer for use with the physiological sensor. The HMM-based recognizer was designed using HTK, a product of the former Entropic Research Laboratories. Like the DTW-based Clamor recognizer, RSC's HMM-based recognizer provided discrete recognition for up to several distinct tokens. The key difference was that with an HMM-based recognizer, additional training samples could be used to re-estimate the speech models, and presumably build a more robust, statistically accurate model of each token as more and more training utterances were collected from the user. The refined HMM models should perform better, while still maintaining the same level of computational complexity. With the DTW approach, the use of additional user utterances for recognizer training would necessarily increase the computational burden of speech recognition at runtime – the more templates that were collected, the longer each match would take.

In order to support rapid integration and testing of user interfaces involving the physiological sensor, it was integrated with Rockwell's Automatic Speech Recognition (ASR) Server technology. The ASR Server provided abstraction of an encapsulated speech recognition engine (Clamor was used for the physiological sensor) through a platform-neutral TCP/IP socket interface. Applications could be quickly designed to exploit speech recognition services of the ASR Server through a simplified protocol. The ASR Server could, in turn, provide speech recognition through either the physiological sensor, or a conventional acoustic microphone. The physiological sensor was demonstrated in conjunction with Rockwell's Multimodal Integrated Displays Testbed in early 1999.

In early 2000, RSC's HMM-based recognizer for the physiological sensor was integrated with RSC's Bimodal ASR Server. The Bimodal ASR Server employed a subset of the same client/server interface protocol used by the ASR Server; whereas the ASR Server encapsulated COTS acoustic speech recognizers, the Bimodal ASR Server encapsulated more experimental recognition technologies, including both the HMM-based recognizer for the physiological sensor, as well as the visual lip-tracking based speech recognizer described in elsewhere in this text. The physiological sensor and Bimodal ASR Server were demonstrated as components of Rockwell's Integrated Displays Testbed v2 in early 2000 [Vassiliou, 00]. As part of the demonstration, a user could dynamically switch between speech recognition using either the lip tracker or the physiological sensor.

The natural extension of this work would be development of a hybrid speech recognition technology that concurrently uses both the physiological sensor and the visual speech recognizer. The two technologies are uniquely complementary because while the visual speech recognizer leverages key visible features of speech articulator motion (vital for recognition of consonant sounds), it is unable to distinguished voiced from unvoiced speech, and indeed is fairly unsuitable for discrimination of vowel sounds from one another. On the other hand, because of its nearly direct coupling to the vocal tract, the physiological sensor is advantageously placed for detecting voicing and discriminating vowel sounds, while its ability to capture subtle acoustic transients of consonant production may be compromised by its body-coupled nature. The visual speech recognizer is already HMM based, so significant research opportunities exist for the development of appropriate feature vectors and HMM topologies to integrate the two distinct signal streams (visual & acoustic).

3.1.3 Ergonomics

The physiological sensor was found to be generally comfortable to wear, though there were some issues with the design. One obvious problem was that users wearing a collared dress shirt could have problems fitting the physiological sensor band either above or under the collar. Generally, with a shirt collar closed, fitting the physiological sensor inside the collar band was not practical. Wearing the physiological sensor higher on the neck than a closed shirt collar tended to limit head movement. Possibly, a narrower band and smaller sensor capsule could help with these issues.

The neck band itself was fairly easy to secure due to the use of Velcro surfaces. The fabric of the neck band was of a dense weave, which could lead to the accumulation of perspiration under the neck band under some conditions. A thinner, more loosely woven fabric, perhaps an elastic one, might be helpful.

The physiological sensor was also compared to a similar COTS throat worn microphone product, the LASH II microphone distributed by Television Equipment Associates. While the LASH II did use a thinner, narrower, elastic collar band, the plastic hook assembly for closing and securing the LASH II was not as easy to use as the Velcro design of the ARL physiological sensor. Further, the LASH II design caused two rigid plastic nodes to be pressed against the user's throat, which could cause significant discomfort when worn over extended periods. In contrast, wearers generally did not find the ARL physiological sensor to increase in discomfort over time.

Some hesitance and psychological resistance to wearing the physiological sensor was also reported of prospective users. An obvious safety concern for any neck worn apparatus is the possibility of choking, either by accident or by assailants. Also, while head worn microphones of some styles have come to be socially acceptable to wearers and even fashionable or "cool" in certain contexts, the visual appearance of the neck worn physiological sensor was less acceptable to some users.

3.1.4 Physiological Sensor Integration Issues

In early 1999, Rockwell received first samples of the ARL Physiological Sensor technology. Early samples used a fairly large (~5"x3"x2") preamplification module, which was rather bulky and not well suited to bodyworn applications. Despite having a full metal casing, the combination of physiological sensor and preamplification module was also susceptible to grounding problems, which would cause a strong 60Hz hum to be present in the output signal. The grounding problems were corrected in the next received prototype early in 1999 and the physiological sensor was successfully mated to a PC-based sound card using the line level input. Some speech recognizers are designed with the assumption that the microphone input of a sound card will be used for speech acquisition, so the user of line level input could have been an integration issue for some speech recognition technologies.

Newer versions of the physiological sensor supplied by ARL in late 1999 and early 2000 used a much smaller and lighter preamplification module (~2"x1"x.5") in a plastic rather than a metal housing. The new preamplification module was light enough to be carried with the user, and the signal level was suitable for use with the microphone inputs of typical PC sound cards. It also included a momentary push-to-talk switch. Conceptually, a push-to-talk switch is helpful in speech recognition applications because if the press and release events for the switch can be detected by the speech recognizer, then delimitation of user utterances becomes fairly easy. Also, the use of a push-to-talk switch helps to prevent false recognition (insertion) errors where extraneous noises or speech not intended for the recognizer are acquired by the transducer. In the case of the current versions of the physiological sensor though, the implementation of the push-to-talk switch is suboptimal for speech recognition. First, the switch is electromechanical and entirely embedded in the preamp module of the physiological sensor, so there is not a deterministic way (e.g. additional connector pin) for an attached device or computer to ascertain when the switch is pressed and released. The switch also induces significant transients in the sensor's output signal when it is pressed and released. Such transients in the speech signal are apt to confuse most existing speech recognition technologies. The workaround solution employed to address these issues was to keep the push-to-talk switch depressed at all times while using a speech recognition system, and to rely on other, external push-to-talk switch mechanisms that were more readily tracked by the Rockwell ASR Server. Additionally, because the push-to-talk switch was of a momentary-on design, additional external fixtures were required to keep the switch depressed.

For some applications, it was desirable for the user of the physiological sensor to be free to move about untethered. Attempts were made to connect the physiological sensor to a wireless microphone transmitter module (Audio Technica ATW-T75), but the output signal levels and impedance were found to be not fully compatible with the input stages available on the wireless transmitter. Although a signal could be sent wirelessly, additional distortions were introduced, which ultimately degraded speech recognition accuracy.

RSC has provided ARL with recommendations for improvements to the design of future Physiological Sensor based microphones.

3.2 Army Research Laboratory

ARL has conducted two experiments using the Physiological Sensor as an input device for ASR. The first effort used the Entropic HTK as the automatic speech recognition (ASR) engine and compared the capabilities of the Physiological Sensor with an acoustic microphone. The second effort utilized Dragon Systems Naturally Speaking, a commercial ASR product to evaluate the possibility of using the Physiological Sensor with commercial speech engines.

All applications of the Physiological Sensor as a speech input device must take into account the difference in frequency response of this sensor as compared to a typical airborne acoustic microphone. This difference in frequency response typically precludes the use of acoustic language models provided with most ASR systems.

3.2.1 Physiological Sensor with Entropic HTK

For the experiment using HTK, ARL teamed with the United States Military Academy (USMA) to develop speech models appropriate for use with the Physiological Sensor [Bass, 99]. The Entropic HTK, a Hidden Markov Model based system, was chosen because it provides the flexibility required to adapt the internal configuration of the ASR engine for use with the Physiological Sensor.

The test consisted of trying to recognize one of 50 phrases using both an airborne sensor (microphone) and the Physiological Sensor. Two recognizers were used, each trained on one of the sensors being tested. Phrases consisted of two to ten words each, with a total of 153 unique words. Each test subject spoke the phrases in an environment that yielded speech to noise ratios of 0-, 3-, and 10dB SNR through the airborne sensor, while wearing both the airborne and physiological sensors.

Speech training and testing was conducted by USMA at their facilities. Training was performed using data collected from 21 subjects speaking the 50 phrases in a quiet environment. The result of the training is a speaker independent model for recognition of the 50 test phrases. Testing was then performed on data collected using 14 new subjects to speaking the 50 phrases in each of the given noise environments.

The results of this experiment are shown in tables 1 and 2. In all cases the Physiological Sensor and related recognizer outperformed the airborne acoustic sensor and related recognizer for the given noise levels. Further, the % accuracy of the Physiological Sensor degrades at a much lower rate with increased noise as compared to the airborne acoustic sensor.

3.2.2 Physiological Sensor with Dragon Naturally Speaking

In order to evaluate other possible application areas for the Physiological Sensor we decided to perform a limited test with a commercial ASR product. We selected Dragon Naturally Speaking for the test because we had considerable experience using this product. To simplify the experiment we used the same set of phrases as used with the HTK testing. One user trained the system using the standard user training session. In addition, all of the words in the command phrases were trained separately.

With this very limited data set, 50 phrases and one user, recognition rates were found to vary between about 60% and 80%. While not outstanding, this is a fairly good result considering that the ASR engine was developed for an airborne acoustic microphone. It should be noted that the worst recognition rates were obtained when the user removed and reattached the Physiological Sensor. We assume that changes in the sensor pressure and position are the cause for these variations. No tests were performed in the presence of noise.

3.2.3 Future Research and Experimentation

Experiments with the Physiological Sensor have demonstrated its capability to be used as a speech sensor for specially trained and configured ASR systems. The requirement for special configurations prevents the application of this sensor with many of the commercial ASR products on the market. Since the private sector is investing heavily in the development of these continually improving commercial ASR products it makes sense to leverage this effort. As a result, ARL will work to develop methods to convert the output of the Physiological Sensor into a signal that more closely approximates that of an acoustic sensor. If we can accomplish this then the Physiological Sensor should be suitable for use with any commercial ASR product. The resulting system would have the improved capabilities of the commercial ASR products with the noise rejection capability of the Physiological Sensor.

4. Summary, Conclusions (Lessons Learned), and Recommendations

Several areas exist to improve the operation of the Physiological Sensor as a microphone. The sensor already has good airborne noise rejection, but more can be done to limit the amount of airborne noise that couples to the sensor. An acoustic insulation material can be incorporated around the shroud of the sensor that contacts the skin to prevent the airborne noise from contacting the sensor's gel pad. Additionally, sensors could be mounted on both sides of the throat and their outputs summed simultaneously so that the speech would add constructively, whereas the noise would be reduced by common mode rejection. Since the vocal folds are not always symmetrical, the combined left and right signal may improve intelligibility through construction of an enhanced signal.

One potential problem in the application of the Physiological Sensor as an input to ASR systems is the substantial variation in signal due to changes in sensor pressure and position. We will research this issue in the future and attempt to minimize these effects in order to improve operation with ASR software.

Circuit modifications can made to eliminate noises from switch activation, match impedance for interaction with other devices, and adjust the filtering to create a more accurate representation of the speech. The preamplifier used in all of the experiments described herein had a flat response, and did not enhance or boost the high frequencies that are lower in amplitude than the very dominant lower formants. Developing a non-linear amplifier (filter) can reduce the "through the wall" perception developed by some listeners, and may produce waveforms that better match what the commercial ASR engines expect. In addition, refinement of ergonomics and packaging would be worthwhile for maturing this technology into a product.

The physiological sensor has demonstrated exceptional capabilities for the detection of voice in high noise environments. In addition, the physiological parameters detected by this sensor provide health and performance indication, but might ultimately provide invaluable emotional or physiological data that can be used to adapt and optimize ASR algorithms under diverse situations. This is important in almost every military and civilian application. Acoustics can provide invaluable clues to help understand the interrelations between the soldier's physiology, the task at hand, the spoken word's intent, and the surrounding environment.

Areas requiring future research include the development of a user independent HMM model set to assist developers working with of the Physiological Sensor, development of algorithms or filters to enhance operation of the sensor for use with commercial ASR products, and refinements in overall operation.

References

[Bass, 99] J. Bass, M. Scanlon, T. Mills, and J. Morgan, “Getting Two Birds with One Phone: An acoustic sensor for both speech recognition and medical monitoring”, *Acoustical Society of America*, November 1999.

[Scanlon, 98] M. Scanlon, “Acoustic Sensor for Health Status Monitoring”, *Proceedings of the 1998 Meeting of the IRIS Specialty Group on Acoustic and Seismic Sensing*, Volume II, pp. 205-222.

[Scanlon, patents] M. Scanlon, “Sudden infant death syndrome (SIDS) monitor and stimulator”, May 1996, U.S. Patent 5,515,865; “Motion and sound monitor and stimulator”, Nov. 4, 1997, U.S. Patent 5,684,460; “Acoustic monitoring sensor”, December 29, 1998, U.S. Patent 5,853,005.

[Vassiliou, 00] M. Vassiliou, V. Sundareswaran, S. Chen, R. Behringer, C. Tam, J. McGee, “Integrated multi-modal human-computer interface and augmented reality for interactive display applications”, *2000 SPIE Aerosense*, April 24-28, 2000, Orlando FL.

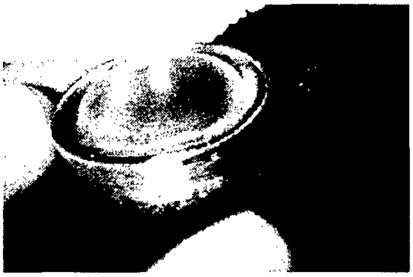

Figure 1: Gel sensor pad.

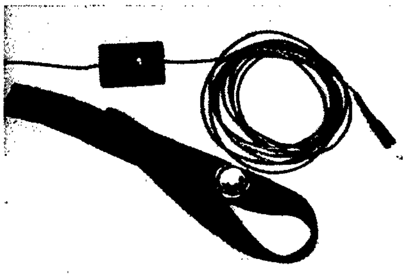

Figure 2: Neck assembly for voice.

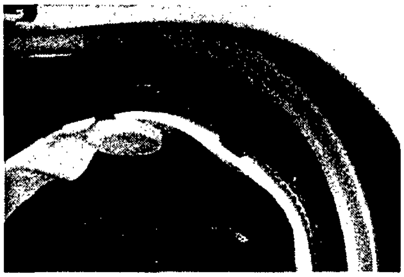

Figure 3: Sensor in helmet headband.

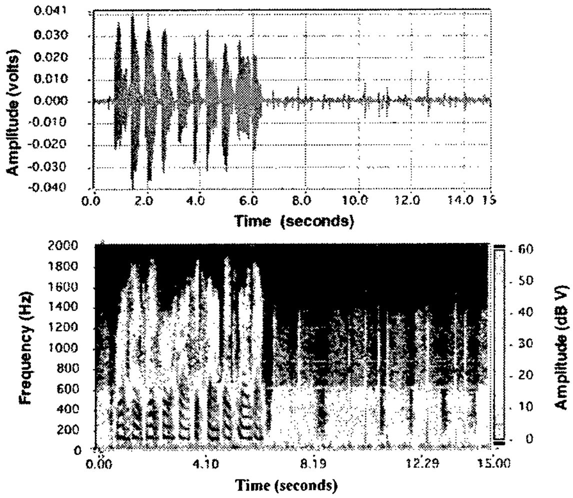

Figure 4: Fluid sensor held at throat for 1 to 10 voice count and mouth breaths.

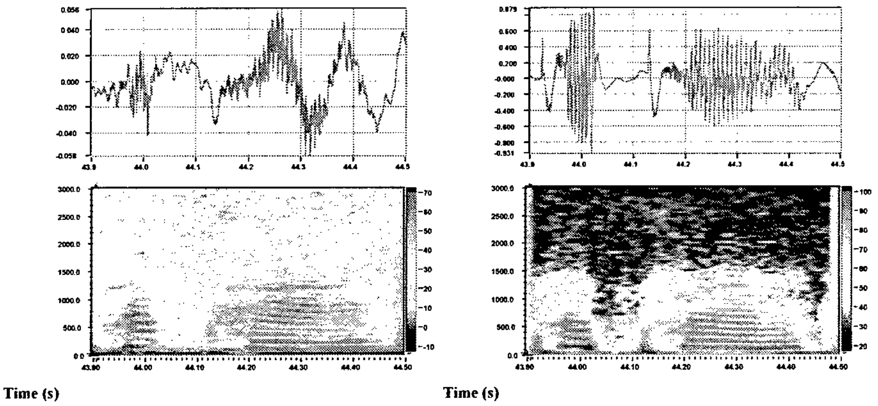

Figure 5: Comparison of spoken word "papa" taken with ambient microphone (left) and throat pad (right).

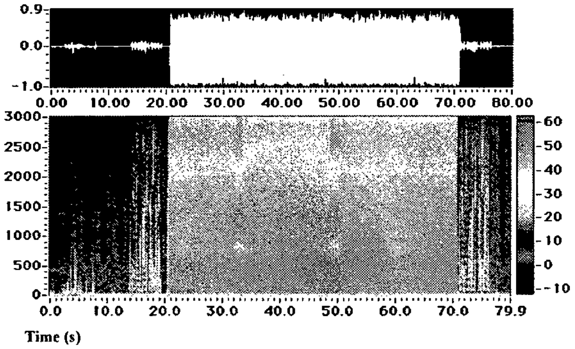

Figure 6: Boom microphone detecting voice.

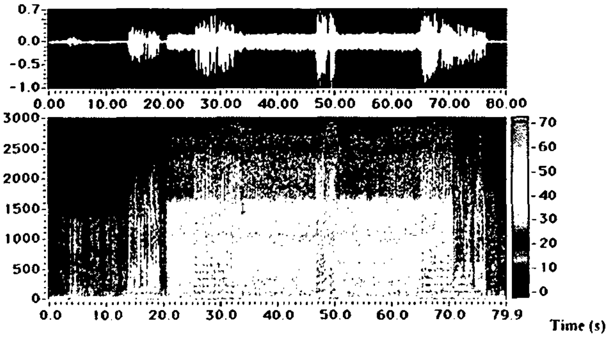

Figure 7: Gel sensor on neck detecting voice.

Table 1. Sentence Loop Language Model

Sentence Loop Model (% Perfect Sentence Recognition)		
dB	Airborne	Physiological
0	40.6	96.5
3	60.7	98.7
10	98.7	99.4

Table 2. Word Loop Language Model

Word Loop Model (% Perfect Sentence Recognition)		
dB	Airborne	Physiological
0	-0.1	39.6
3	12.8	50.5
10	51.5	66.8